Patient journey foundational model for scalable imputation of missing units of measurement (UoM) in electronic health records data

Ehsan Alipour, Wilson Lau, Youngwon Kim, Sihang Zeng, Anand Oka, Jay Nanduri

Poster # RWD140

Background

- Missing units of measurement (UoM) for labs, observations, and measurements create major data quality challenges in EHR data.
- Missing UoM can complicate clinical interpretation, quality control, and research use of data.
- In some EHR instances, UoM missingness can reach up to 50%.
- Existing rule-based approaches often miss clinical context such as history, age, and sex.

Methods

- Patient journeys were constructed from three parallel sequences: event name, event value, and patient age at event.
- Sequences were tokenized using a custom tokenizer and passed through a generative pre-trained transformer (GPT) model with 150 million parameters, augmented with a prediction head for UoM for each input token.
- The model was trained on two tasks:

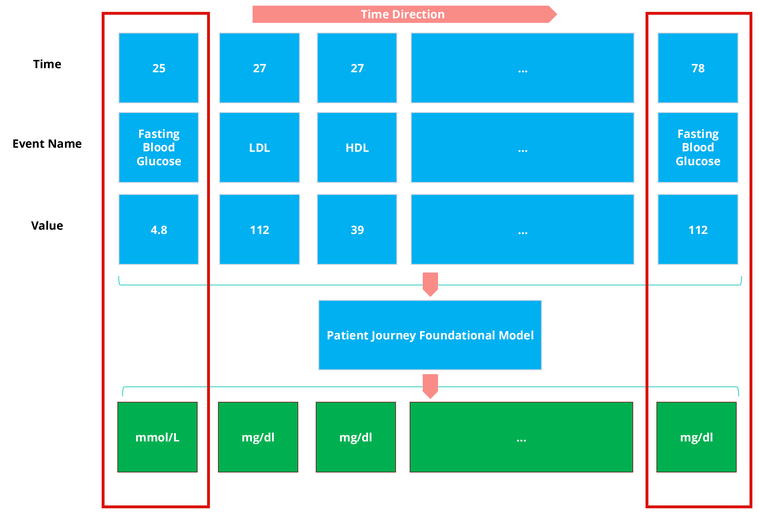
1) Prediction of next token for each token position including event name, time and value
2) Prediction of UoM for each event in the patient journey

- Events without UoM, such as diagnosis codes, defaulted to Not Applicable (N/A).
- Pretraining enabled the model to learn longitudinal patient context, including age and medical history, to support UoM prediction.

Objective

To develop and evaluate a transformer-based patient journey foundation model (PJFM) for imputing missing units of measurement (UoMs) using longitudinal patient context.

Patient journey model architecture



The model uses longitudinal patient context to predict UoM at the token level.

Model performance

UCUM Unit Label	UCUM Name	Precision	Recall	F1-score
mg/dL	milligram per deciliter	0.997	0.998	0.997
%	percent	0.991	0.992	0.992
mmol/L	millimole per liter	0.979	0.989	0.984
10^3/uL	thousand per microliter	0.975	0.982	0.978
g/dL	gram per deciliter	0.989	0.992	0.99

Performance for top units of measurement

Key result

98% accuracy imputing UoM across **438 unique units**; weighted F1-score **0.90**.

Results

- Identified 438 unique units of measurement (UoMs) in the dataset.
- Achieved strong UoM imputation performance: 98% accuracy, 0.90 weighted F1-score, 0.90 precision, and 0.89 recall.
- Performance was strongest for frequently occurring UoMs.
- For common measurements, such as body weight and height, predictions aligned with realistic BMI interpretations, suggesting context-aware unit assignment.
- The model also performed well on lab tests with multiple valid units.
- For C-reactive protein (CRP), which can be reported in mg/dL or mg/L with overlapping value ranges, the model achieved 94% and 89% accuracy, respectively.
- Results suggest the model can infer UoMs even when numerical values alone are insufficient.

Clinical measurement examples

Measurement	UCUM Unit Label	UCUM Name	Accuracy (%)
Body weight	g	gram	100
Body weight	kg	kilogram	99
Body weight	[oz_av]	ounce (avoirdupois)	99
Body weight	[lb_av]	pound (avoirdupois)	99
Body height	cm	centimeter	99
Body height	[ft_us]	foot (US)	100
Body height	[in_us]	inch (US)	95
Body height	[in_i]	inch (international)	99
Body height	m	meter	83
Body temperature	Cel	degree Celsius	100
Body temperature	[degF]	degree Fahrenheit	100

Examples include height, weight, and temperature units.

Discussion

- The patient journey foundation model (PJFM) demonstrated strong capability in imputing missing UoMs by learning contextual patterns from longitudinal patient journeys.
- Unlike rule-based approaches that rely on predefined mappings, the model uses sequence-level context to resolve ambiguous units.
- This context-aware approach supports more flexible and generalizable UoM inference across diverse EHR data.
- The model can assign appropriate units even when multiple valid UoMs have overlapping numeric ranges, a common challenge in clinical data.
- By incorporating broader patient information, the PJFM may provide a more consistent and scalable approach to handling missing UoMs.

A patient journey foundation model can accurately impute missing units of measurement, improving EHR data completeness for downstream clinical and research use.